CONCEPTUAL APPROACH TO SURVIVAL ANALYSIS

Joanna H. Shih, Ph.D.

OBJECTIVES

1. To describe the basic features of survival data.
2. To introduce the Kaplan-Meier survival curve.
3. To present the approaches for comparing two survival curves.
4. To provide motivation and summary of stratified analysis and Cox's regression analysis for survival data.

OUTLINE

1. Basic Features of Survival Data
    - Time to event data
    - Censoring
2. Examples
3. Survival Curve
    - Motivation
    - Definition and graphical presentation
    - Statistical inference
    - Example
4. Comparison of Two Survival Curves
    - Point-by-point comparison
    - Comparing two survival curves - logrank test
    - Example

5. Other Topics

   - Stratified logrank test
   - Cox's regression model
   - Example

REFERENCES

1. Altman, D.G. (1991). *Practical Statistics for Medical Research.* Chapman & Hall. New York.

2. Fisher, L.D. and Van Belle. (1993). *Biostatistics: A Methodology for the Health Sciences.* John Wiley & Sons. New York.

3. Friedman, L., Furberg, C.D. and DeMets, D.L. (1996). *Fundamentals of Clinical Trials.* Mosby-Year Book.

4. Kalbfleisch, J. and Prentice, R. (1980). *The Statistical Analysis of Failure Time Data.* John Wiley & Sons. New York.

Basic Features of Survival Data

1. Main interest focuses on the time taken for some event to occur. **Survival time** is the time from some fixed starting point to the onset of the event.

    - In controlled clinical trials, time of entry to a study and time of the event (death, MI, stroke, etc.) are recorded.
    - In laboratory studies, often the starting time point is the same for all the subjects and time of the event is recorded.

2. The event of interest is almost never observed in all subjects. The survival time is **censored** if the event is not observed at the end of the study, to indicate the period of observation was cut off before the event occurred.

    - For example, in a study to compare the survival of patients having different types of treatment strategies for atrial fibrillation, although the patients will be followed up for several years there will be many who are still alive at the end of the study. The exact survival times for those patients are unknown; it is only known that their survival times will be longer than their times in the study.
    - Other reasons of censoring: withdrawal, leaving the study because of moving to a different area. Censoring due to reasons unrelated to the outcome of the study is called **independent censoring**.

Examples

1. A hypothetical example. It illustrates the different ways in which patients can proceed through a study. Patients are recruited during a six month period and then followed up for a minimum of 12 months. Thus the patients are observed for between 12 and 18 months, the earliest accrued patients being observed for the longest time.

2. A clinical trial investigating the effect of prednisolone for patients suffering from chronic active hepatitis (Kirk, A.P. *et al.*, 1980).

   - Forty-four patients with chronic active hepatitis were randomized to either prednisolone (n = 22) or an untreated control group (n = 22).
   - Outcome: Survival times and vital status.

3. A nonrandomized study comparing the immunotherapies BCG (Bacillus Calmette-Geurin) and *c. parvum* (*corynebacterium parvum*) for their abilities to prolong remission and survival times for melanoma patients.

   - Thirty melanoma patients receiving either BCG or *c parvum* were resected before the treatment began.
   - Prognostic factors: age, sex, disease stage.
   - Outcome: disease free survival (survival without relapse). survival time = minimum of time to relapse and time to death.

Survival Curve

1. Motivation: To understand the survival experience of a population in various time points.

2. Definition: The graphical presentation of the total survival experience during the period of observation is called survival curve, and the mathematical presentation is called survival function.

   Survival function $S(t)$ = probability of surviving beyond time t.

   $$0 \leq S(t) \leq 1.$$

3. **Kaplan-Meier** survival curve is the estimate of the survival function from the sample available.

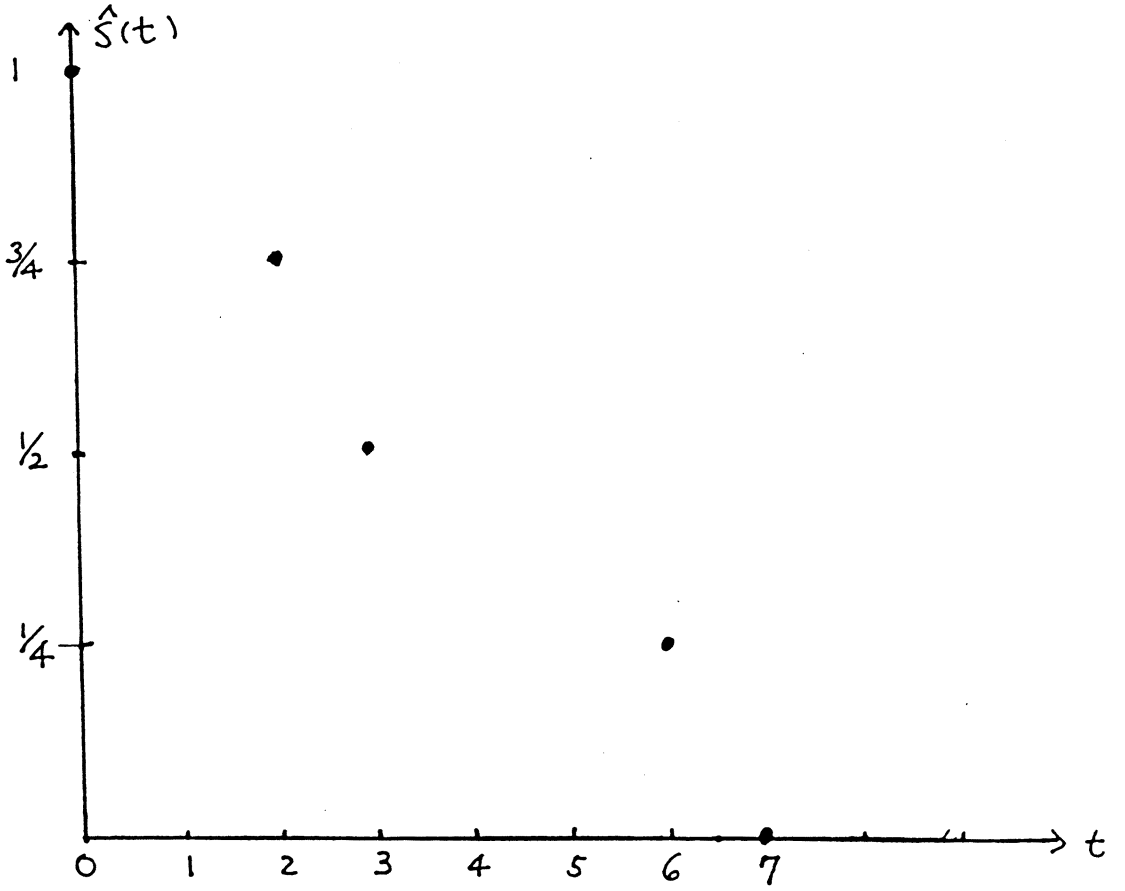
$$\begin{aligned}
\hat{S}(t) &= \text{estimated percentage surviving beyond time } t \\
&= \left(\frac{\text{no. surviving beyond } t}{\text{no. surviving beyond } t + \text{no. dead at } t}\right) \times \\
&\quad \text{estimated percentage surviving up to time } t \\
&= \left(1 - \frac{\text{no. dead at } t}{\text{no. surviving beyond } t + \text{no. dead at } t}\right) \times \\
&\quad \text{estimated percentage surviving up to time } t
\end{aligned}$$

4. Illustration

Case I.

Subject no. (k)	Survival time (t_k)	no. at risk (r_k)	no. dead (f_k)	$1 - \frac{f_k}{r_k}$	$\hat{S}(t_k)$
1	2	4	1	3/4	3/4
2	3	3	1	2/3	$2/3 \times 3/4 = 1/2$
3	6	2	1	1/2	$1/2 \times 1/2 = 1/4$
4	7	1	1	0	$0 \times 1/4 = 0$

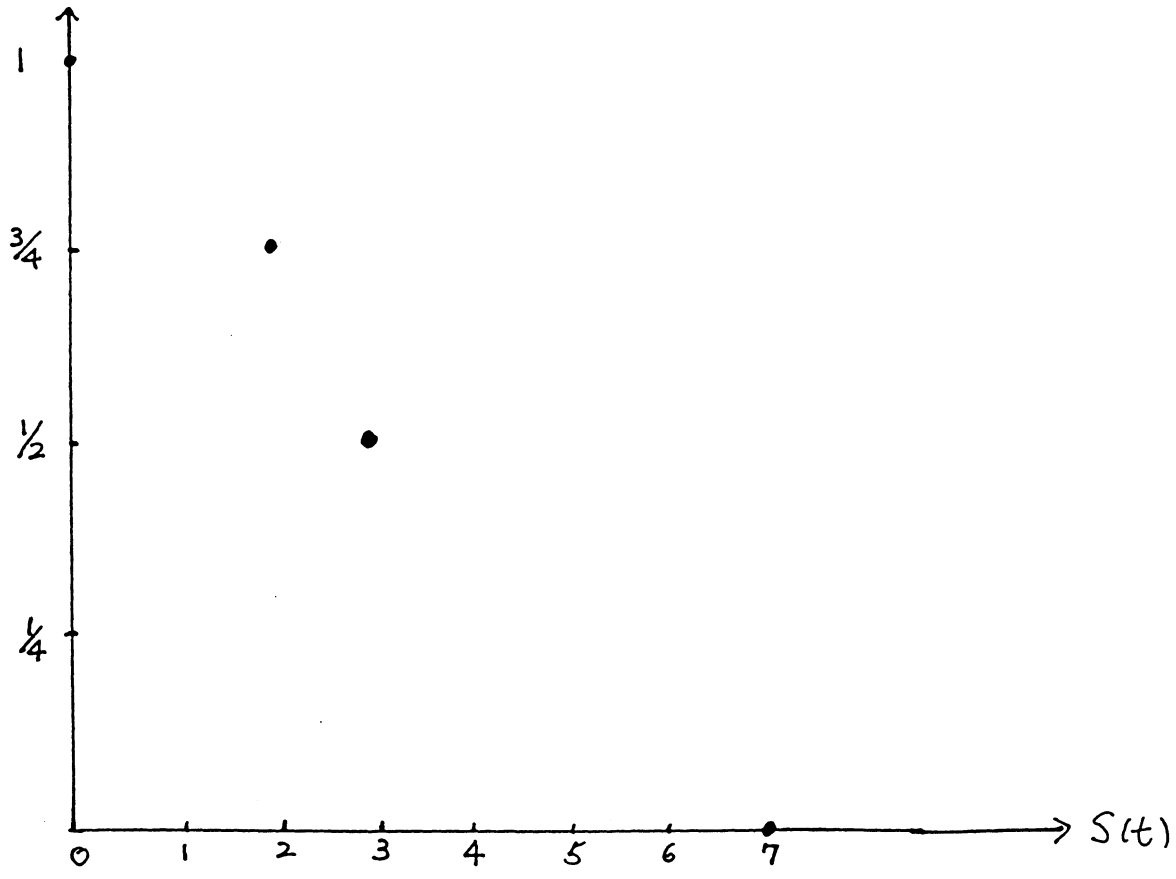
Kaplan-Meier survival curve:



Case II.

Subject no no. (k)	Survival time (t_k)	no. at risk (r_k)	no. dead (f_k)	$1 - \frac{f_k}{r_k}$	$\hat{S}(t_k)$
1	2	4	1	3/4	3/4
2	3	3	1	2/3	$2/3 \times 3/4 = 1/2$
3	6+	2	0	1	$1 \times 1/2 = 1/2$
4	7	1	1	0	$0 \times 1/2 = 0$

Kaplan-Meier survival curve:



5. Statistical inference:

   - Estimate the probability of survival at a given time t, $\hat{S}(t)$.
   - Estimate the variability of $\hat{S}(t)$. The variance of $\hat{S}(t)$ is asymptotically equivalent to

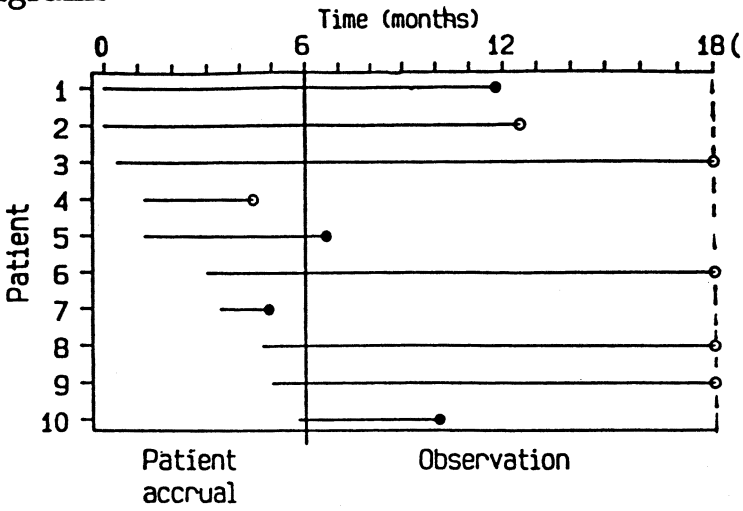
$$\mathrm{Var}(\hat{S}(t)) \approx \hat{S}(t)^2 \sum_{k:t_k \leq t} \frac{f_k}{r_k(r_k - f_k)}.$$

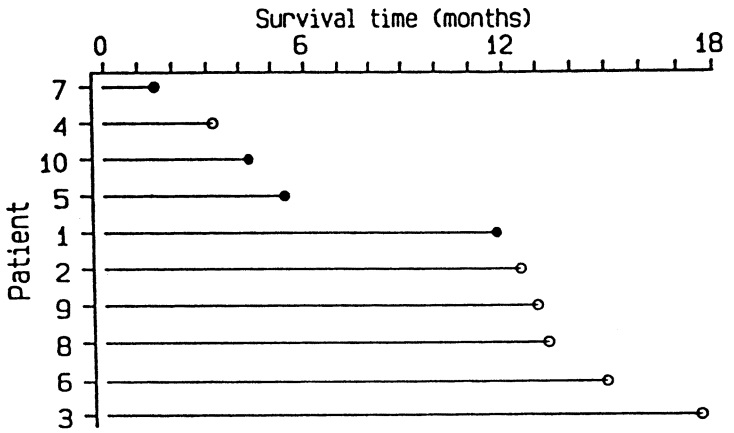
6. Example 1.

   - Data.

Patient	Time at entry (months)	Time at death or censoring (months)	Dead or censored	Survival time
1	0.0	11.8	D	11.8
2	0.0	12.5	C	12.5
3	0.4	18.0	C	17.6
4	1.2	4.4	C	3.2
5	1.2	6.6	D	5.4
6	3.0	18.0	C	15.0
7	3.4	4.9	D	1.5
8	4.7	18.0	C	13.3
9	5.0	18.0	C	13.0
10	5.8	10.1	D	4.3

- Diagrams

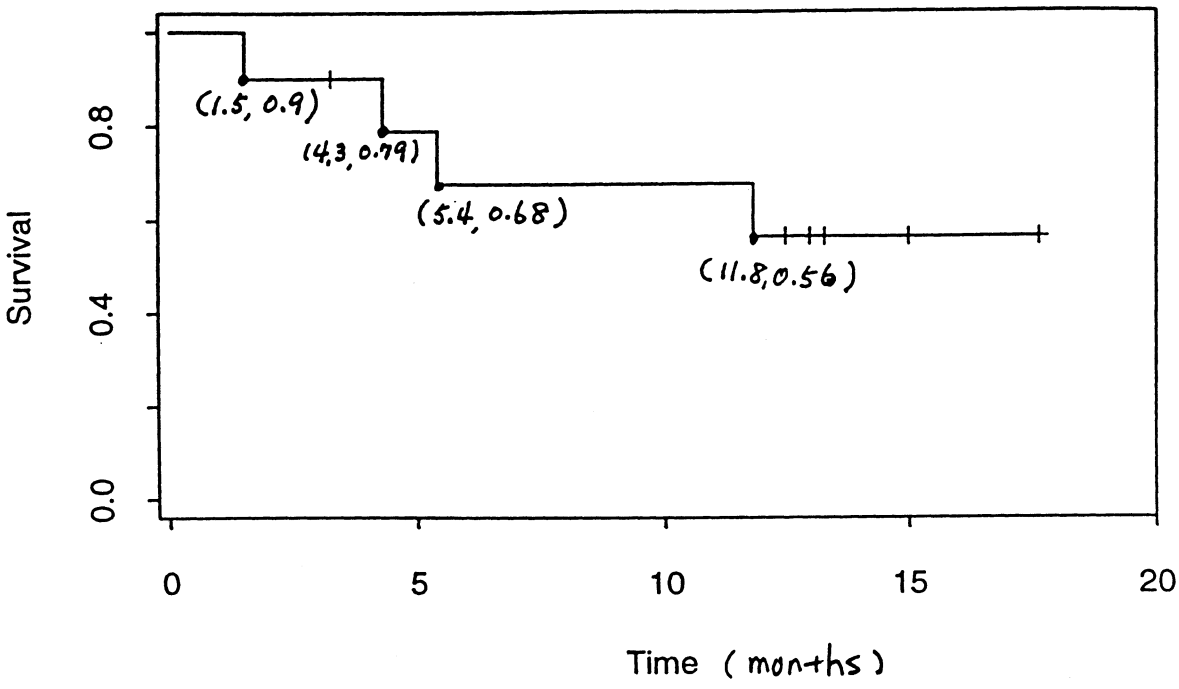




• Dead o censored

- Software: SAS, BMDP, SPSS, SPLUS.

- Kaplan-Meier curve.



Comparison of Two Survival Curves

1. Comparing two survival curves at a given time point, t.

   - Forming a statistical hypothesis.

$$\begin{aligned} H_0 &: S_1(t) = S_2(t) \\ H_a &: S_1(t) \neq S_2(t) \quad \text{two-sided} \\ H_a &: S_1(t) > S_2(t) \quad \text{one-sided} \\ H_a &: S_1(t) < S_2(t) \quad \text{one-sided} \end{aligned}$$

   - Calculating the Kaplan-Meier survival estimate at time t for each sample, $\hat{S}_1(t), \hat{S}_2(t)$.
   - Specifying the type I error (α).
   - Calculating the test statistic,

$$Z = \frac{\hat{S}_1(t) - \hat{S}_2(t)}{\text{SE}(\hat{S}_1(t) - \hat{S}_2(t))},$$

     where

$$\text{SE}(\hat{S}_1(t) - \hat{S}_2(t)) = \sqrt{\text{var}(\hat{S}_1(t)) + \text{var}(\hat{S}_2(t))}.$$

     Z approximately has a standard normal distribution under H_0 - reference distribution.

   - If Z is in the upper or lower $100 \times \alpha/2\%$ ($100 \times \alpha\%$ for one-sided test) of the reference distribution, then we reject H_0.
   - Calculating the $p-$value, the probability of observing a Z-value more extreme than the one from the current sample if the null hypothesis is true.

2. Comparing two survival curves.

- Comparing the whole curves rather than a point.
- Logrank test (Mantel-Haenszel test).
  - Arrange the distinct survival times from the two groups in an ascending order, excluding censored survival times: $\{t_1, t_2, \cdots, t_K\}$.
  - At each time t_j, construct a 2×2 table

	No. dead	No. surviving	Total
Group 1	a_j	b_j	$a_j + b_j$
Group 2	c_j	d_j	$c_j + d_j$
Total	$a_j + c_j$	$b_j + d_j$	n_j

If the null hypothesis is true, then the expected no. of deaths at group 1, denoted by $\mathrm{E}(a_j)$, is equal to

$$E(a_j) = (a_j + b_j)(a_j + c_j)/n_j,$$

$$Var(a_j) = \frac{(a_j + b_j)(a_j + c_j)(b_j + d_j)(c_j + d_j)}{(n_j - 1)n_j^2}$$

  - Form K 2×2 tables, and calculate the test statistic using the results from these tables.

$$Z = \frac{\sum_{j=1}^{K} a_j - \mathrm{E}(a_j)}{\sqrt{\sum_{j=1}^{K} \mathrm{Var}(a_j)}}.$$

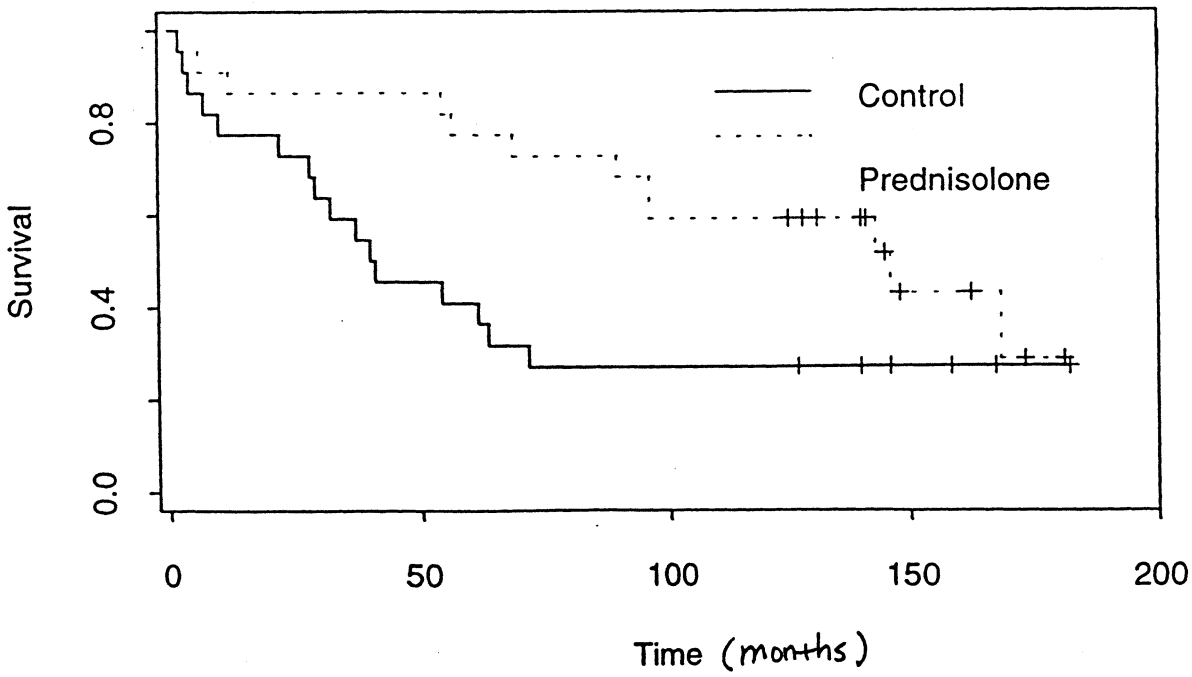
  - Z approximately has a standard normal distribution under H_0.
  - If Z is in the upper or lower $100 \times \alpha/2\%$ ($100 \times \alpha\%$ for one-sided test) of the reference distribution, then we reject H_0.
  - Calculating the $p-$value.

3. Example 2 - patients with hepatitis.

- Data.

Control survival times	Prednisolone survival times	Control survival times	Prednisolone survival times
2	2	41	131+
3	6	54	140+
4	12	61	141+
7	54	63	143
10	56	71	145+
22	68	127+	146
28	89	140+	148+
29	96	146+	162+
32	96	158+	168
37	125+	167+	173+
40	128+	182+	181+

- Calculating the Kaplan-Meier survival curve for each



- Comparing the survival difference at 5 years (60 months).

  $H_0 : S_c(60) = S_p(60) \qquad H_a : S_c(60) < S_p(60).$

  $\alpha = .05.$

  $\hat{S}_c(60) = .4090 \qquad \hat{S}_p(60) = .7727.$

  $\text{SE}(\hat{S}_c(60) - \hat{S}_p(60)) = .1377.$

  $Z = -2.64$. Compare Z with the lower 5th percentile of the standard normal distribution (= -1.645). Reject the null hypothesis because $Z < -1.645$.

  p-value $< .01 < \alpha$.

- Performing the log-rank test.

  $H_0 : S_c(.) = S_p(.), \qquad H_a : S_c(.) < S_p(.).$

  $Z = 2.0$. Compare Z with the upper 5th percentile of the standard normal distribution (1.645). Reject the null hypothesis because $Z > 1.645$.

  p-value $< \alpha$.

Other Topics

1. Stratified analysis.

   - Motivation
     - If important prognostic factors (covariates) are imbalanced at the study entry between the two groups, the survival analysis may be influenced by the difference observed in the prognostic factors.

Intervention	Control
Young	Old

       The difference observed in the intervention and control groups may be due to the age difference. In this case, the intervention effect and age effect are confounded.
     - The analysis may be biased.
     - In large randomized clinical trials, prognostic factors are often balanced. But in nonrandomized studies or studies of moderate size, the balance may not be assured.
   - Stratified logrank test
     - Divide the data according to the levels of the significant prognostic factors (e.g. race, age group, etc.).
     - Calculate the logrank test within each stratum and accumulate the results over strata.

       a_{ij} = no. of deaths at time t_j in the ith stratum.

       $$Z = \frac{\sum_i \sum_j a_{ij} - E(a_{ij})}{\sqrt{\sum_i \sum_j \mathrm{Var}(a_{ij})}}.$$

       Z approximately has a standard normal distribution.

2. Regression analysis

- Motivation
  - If there are many prognostic factors, each with several levels, the number of strata can quickly become large with few patients in each stratum, resulting in the loss of power in stratified analysis.
  - If a covariate is continuous, it must be grouped into intervals before it can be used in stratified analysis.
- Cox's proportional hazards model
  - It allows for analysis of survival data adjusting for continuous and discrete covariates.
  - Main assumption: proportional hazards model.
    The **hazard** function, denoted by $h(t)$, represents the risk of having an event in a very short time interval after surviving a given time t.
    Proportional hazard means that the change in a covariate (e.g. blood pressure) results in a proportional change of the hazard in a log scale. Mathematically the proportional hazards model is represented by

$$h(t) = h_0(t) \exp(\beta_1 \times x_1 + \cdots + \beta_p \times x_p),$$

    where $h_0(t)$ is called the baseline hazard, β_k, x_k, $k = 1, \cdots, p$ are regression coefficients and covariates.
  - It allows for the measure of the effect of a covariate on the hazard expressed by the regression coefficient.

- Example: Consider only one covariate, systolic blood pressure(BP).

$$h(t) = h_0(t)\exp(\beta_1 \times \text{BP}).$$

Subject	BP	hazard at time t	log(hazard)
A	110	$h_0(t)\exp(\beta_1 \times 110)$	$\log h_0(t) + \beta_1 \times 110$
B	130	$h_0(t)\exp(\beta_1 \times 130)$	$\log h_0(t) + \beta_1 \times 130$

The change of log hazard $= \beta_1 \times (130 - 110)$ which is the change of the blood pressure times β_1 and does not depend on t.

- The proportional hazards assumption may be violated.
- Estimation of the regression coefficients is complex but is available in most statistical software (SAS, BMDP, SPSS, SPLUS). ome software provides diagnostics for checking the assumption.
- The estimate of each regression coefficients $\hat{\beta}$ has approximately a normal distribution. The Z statistic

$$Z = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

has a standard normal distribution under the null hypothesis $H_0 : \beta = 0$ implying no covariate effect.

3. Example 3.

   Treatment: immunotherapies BCG (1) or *c. parvum* (0).

   Covariates: age at the entry of the study (year), gender (1: male, 0:female), disease stage (2-4).

Patient	Age	Gender	Disease stage	Treatment received	Survival times (months)
1	59	0	3	1	33.7+
2	50	0	3	1	3.8
3	76	1	3	1	6.3
4	66	0	3	1	2.3
5	33	1	3	1	6.4
6	23	0	3	1	23.8+
7	40	0	3	1	1.8
8	34	1	3	1	5.5
9	34	1	3	1	16.6+
10	38	0	2	1	33.7+
11	54	0	2	1	17.1+
12	49	1	3	0	4.3
13	35	1	3	0	26.9+
14	22	1	3	0	21.4+
15	30	1	3	0	18.1+
16	26	0	3	0	5.8
17	27	1	3	0	3.0
18	45	0	3	0	11.0+
19	76	0	3	0	22.1
20	48	1	3	0	23.0+
21	91	1	4	0	6.8
22	82	0	4	0	10.8+
23	50	0	4	0	2.8
24	40	1	4	0	9.2
25	34	1	3	0	15.9
26	38	1	4	0	4.5
27	50	1	2	0	9.2
28	53	0	2	0	8.2+
29	48	0	2	0	8.2+
30	40	0	2	0	7.8+

- Estimation results.

	$\hat{\beta}$	SE($\hat{\beta}$)	Z
Age	.0102	.0129	.788
Gender	.3317	.5700	.582
Disease stage	1.4115	1.0673	1.323
Treatment	.3240	.5357	.605

$H_0 : \beta_k = 0,\ k = 1, \cdots, 4, \qquad H_a : \beta_k \neq 0.$

$\alpha = .05.$

None of the covariates are significant when comparin $|Z|$ values to the upper 2.5 percentiles of the standard mal distribution ($= 1.96$).

Summary

- We described the basic features of survival data
  - Time to event data
  - Censoring
  - Kaplan-Meier survival curve
- We introduced approaches for comparing the two survival curves
  - point by point comparison
  - whole curve comparison - logrank test
- Other topics
  - Stratified logrank test
  - Cox's regression analysis